KERTAS •AKSES TERBUKA

Analisis Stemming Teks Berita Berbahasa Indonesia dengan Algoritma Porter

Mengutip artikel ini: A. Arif siswandi*dkk*2021*J.Fis.: Conf. Ser.*__1845__012019

Jurnal Fisika: Seri Konferensi **1845**(2021) 012019 doi:10.1088/1742-6596/1845/1/012019

Analisis Stemming Teks Berita Berbahasa Indonesia dengan Algoritma Porter

Arif siswandi[1], A.Yudi Permana[2], Arvita Emarilis[3]

Universitas Pelita Bangsa,Bekasi,Indonesia

Email: arif.siswandi@pelitabangsa.ac.id[1], yudi@pelitabangsa.ac.id[2], arvita@pelitabangsa.ac.id[3]

Abstrak. Stemming adalah proses mengelompokkan berbagai variasi morfologi suatu kata atau kalimat ke dalam satu bentuk dasar yang sama. Dalam stemming bahasa Indonesia, terdapat dua jenis metode stemming yang sudah ada, yaitu algoritma stemming berbasis kamus dan algoritma stemming berbasis non kamus. Pada penelitian ini algoritma yang digunakan adalah algoritma Porter Bahasa Indonesia yang berbasis kamus. Pengujian dilakukan dengan menggunakan 100 dokumen teks bahasa Indonesia yang telah ditentukan. Hasil pengujian yang dilakukan menunjukkan bahwa nilai akurasi tertinggi terdapat pada algoritma Porter, nilai Overstemming dan Understemming terkecil juga terdapat pada algoritma Porter.

Kata kunci: Pengambilan Informasi, Stemming, Akurasi, Overstemming dan Understemming.

1. Perkenalan

Dalam melakukan pencarian informasi terhadap suatu dokumen tekstual atau dikenal dengan Information Retrieval (IR) merupakan suatu proses pemisahan dokumen yang dianggap relevan dari kumpulan dokumen yang tersedia. Dengan bertambahnya jumlah dokumen teks yang dapat diakses di Internet diikuti dengan meningkatnya kebutuhan pengguna akan perangkat pencarian dan informasi yang efektif dan efisien [1]. Efektif berarti pengguna mendapatkan dokumen yang relevan dengan kueri yang dimasukkan. Efisien juga waktu hasil pencarian lebih singkat.

Stemming merupakan proses memetakan varian bentuk kata menjadi akar kata [2]. Stemming merupakan inti dari suatu pengambilan informasi yang efektif dan efisien serta dapat diterima secara luas oleh pengguna. Stemming adalah proses menemukan akar kata suatu kata. Dengan menghilangkan semua prefiks, sisipan, sufiks, dan konfiks (gabungan prefiks dan sufiks) yang baik pada kata turunan. Stemming juga dapat digunakan untuk mendukung proses kategorisasi atau klasifikasi dan clustering. Stemming digunakan untuk mengganti bentuk atau konfigurasi suatu kata menjadi akar kata dan kata tersebut telah sesuai dengan struktur morfologi bahasa Indonesia yang benar dan tepat.

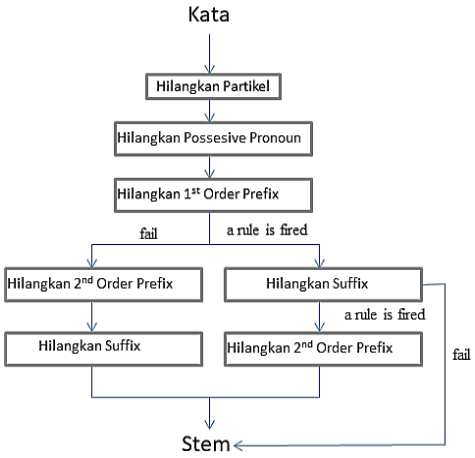
Penggunaan stemming dalam bahasa Indonesia meliputi dua jenis metode stemming yang dikenal, yaitu yaitu kamus berbasis stemming (dictionary based) dan non-kamus berbasis stemming (murni rule based). Pada algoritma stemming yang tidak menggunakan kamus mempunyai margin of error yang relatif tinggi, namun di sisi lain algoritma tersebut mempunyai keunggulan dalam waktu proses yang lebih singkat dibandingkan algoritma stemming yang berbasis kamus. Kebanyakan algoritma stemming masih mengandalkan atau pada kamus untuk memeriksa apakah akar kata suatu kalimat atau kata yang telah dilakukan proses stemming ditemukan atau tidak. Ketika akar kata berhasil ditemukan dalam kamus, proses stemming dihentikan.

Dalam stemming dalam bahasa Indonesia ada beberapa pendekatan seperti stemming Porter, Tala,Vega,

Arifin dan Setiono, Nazief dan Adriani. Hampir tidak ada konsensus umum mengenai efektivitas teknik pendekatan tersebut. Permasalahan stemming lainnya masih bergantung pada beberapa teknik stemming tersebut pada kamus yang lebih luas (comprehensive Dictionary). Pada penelitian akan dilakukan pengukuran efektif terhadap algoritma yang digunakan dalam stemming yaitu kamus menggunakan algoritma porter. Penilaian pengukuran dilakukan pada tingkat akurat, overstemming dan understemming.

2. Algoritma Stemming Porter

Stemming khusus bahasa Inggris ditemukan oleh Martin Porter pada tahun 1980. Suatu algoritma pencarian akar kata diulang dengan melemparkan awalan (atau lebih tepatnya sufiks) pada kata-bahasa Inggris karena dalam bahasa Inggris tidak ada awalan. Karena bahasa Inggris berasal dari kelas yang berbeda, maka dilakukan beberapa modifikasi untuk membuat algoritma porter agar dapat digunakan sesuai dengan bahasa Indonesia. Porter Stemmer untuk bahasa Indonesia berdasarkan bahasa Inggris Porter Stemmer yang dikembangkan oleh WB Frakes pada tahun 1992.



Gambar 1. Alur Proses Stemming Porter Indonesia

Ada lima perangkat aturan pada Algoritma Porter untuk Bahasa Indonesia [1]. Aturannya dapat dilihat pada Tabel 1 sampai dengan Tabel 5.

Tabel 1. Aturan untuk Partikel Infleksional

Sufiks	Penggantian	Ukuran Kondisi	Tambahan Kondisi	Contoh
kah	Batal	2	Batal	sakitkah → sakit
lah	Batal	2	Batal	adalah → ada
permainan kata-kata	Batal	2	Batal	jalanpun → jalan

Meja 2. Aturan Infleksional Kata Ganti Kepunyaan

Sufiks	Penggantian	Ukuran Kondisi	Tambahan Kondisi	Contoh
ku	Batal	2	Batal	Milik → tangan
mu	Batal	2	Batal	tanganmu → tangan
itu	Batal	2	Batal	tangan→ tangan

Tabel 3. Aturan untuk Awalan Turunan Orde Pertama

Awalan	Penggantian	Ukuran Kondisi	Tambahan Kondisi	Contoh
meng	Batal	2	Batal	mengambil → ambil

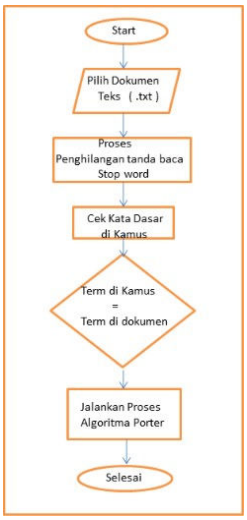
Awalan	Penggantian	Ukuran Kondisi	Tambahan Kondisi	Contoh
tidak	S	2	V...*	menyapa → sapa
laki-laki	Batal	2	Batal	mendapat → dapat
mem	P	2	V...	memilih → pilih
mem	Batal	2	Batal	membeli → beli
Saya	Batal	2	Batal	merusak → rusak
peng	Batal	2	Batal	penguji → uji
sen	S	2	V...	penyayang → sayang
pena	Batal	2	Batal	penduga → duga
pem	P	2	V...	pemikir → pikir
pem	Batal	2	Batal	pembaca → baca
di	Batal	2	Batal	diuji → uji
ter	Batal	2	Batal	tersapu → sapu
ke	Batal	2	Batal	kekasih → kasih

Tabel 4. Aturan untuk Awalan Turunan Orde Kedua

Awalan	Penggantian	Ukuran Kondisi	Tambahan Kondisi	Contoh
ber	Batal	2	Batal	berjalan → jalan
bel	Batal	2	terbuka sedikit	belajar → terbuka
menjadi	Batal	2	Oke...	bekerja → kerja
per	Batal	2	Batal	pertajam → tajam
kulit	Batal	2	terbuka sedikit	pelajar → terbuka
pe	Batal	2	Batal	pekerja → kerja

Tabel 5. Aturan Sufiks Derivasi

Sufiks	Penggantian	Ukuran Kondisi	Tambahan Kondisi	Contoh
kan	Batal	2	Awalan ≠ {ke,peng}	tarikkan → tarik (meng)ambilkan → ambil
sebuah	Batal	2	awalan ≠ {di,meng,ter}	minuman → minum (per)janjian → janji
Saya	Batal	2	awalan ≠ {ber,ke,peng}	(men)dapati → dapat tandai → tanda



Gambar 2. Diagram alir proses stemming Porter Bahasa Indonesia

3. Metodologi Penelitian

Informasi mengenai metode penelitian yang dilakukan dalam penelitian ini. formasi tentang metode penelitian yang dilakukan dalam penelitian ini. Pertama dari objek penelitian, desain penelitian dan teknik pengumpulan data dilakukan dari berbagai sumber, kemudian melakukan preprocessing data dan penerapan algoritma. Langkah selanjutnya yang akan dilakukan adalah melakukan eksperimen atau pengujian data set pada setiap algoritma yang digunakan. Alur proses Stemming Porter bahasa Indonesia tercermin pada gambar proses stemming berikut ini.

Gambar 3. Diagram Alir Proses Stemming Porter Bahasa Indonesia

4. Hasil

Pembahasan akan difokuskan pada hasil uji coba dari kumpulan dokumen yang digunakan, leksikon primer, dan penilaian untuk membandingkan akar kata yang berasal dari akar kata menurut pengetahuan manusia. Berdasarkan penelitian yang telah dilakukan mendapatkan hasil yang berbeda antara algoritma yang satu dengan algoritma yang lain. Pengujian ini dijalankan masing-masing 100 kali untuk setiap algoritma. Untuk mengetahui tingkat kinerja masing-masing algoritma akan dilakukan pengukuran nilai keakuratan kata, yaitu kata overstemming dan understemming.

4.1. Pengumpulan Dokumen

Untuk kumpulan dokumen yang digunakan dalam pengujian, terdapat contoh dokumen sebanyak 100 dokumen, yang sudah dibuat dengan menggunakan ekstensi .txt. Kata-kata dalam dokumen ini diperoleh dari isi artikel dan berita, baik berita maupun artikel di bidang teknik, kesehatan, sains, dan media elektronik. Jumlah kata pada 100 dokumen adalah 25.819 kata. Isi kata pada setiap dokumen bervariasi, mulai dari puluhan kata hingga ribuan kata, dimana kata-kata pada dokumen tersebut belum dilakukan proses pengolahan atau preprocessing.

4.2. Kamus

Semakin lengkap kosakata dasar yang digunakan maka semakin besar nilai akurasi stemmingnya. Dalam penelitian ini, kamus dasar diambil dari daftar default kata utama dalam kamus bahasa Indonesia (KBBI) yang memikat CHM V1.5 yang diunduh dari ebsoft.web.id. Jumlah akar kata dalam kamus adalah 31.295 akar kata [4].

4.3. Penilaian Relevansi

Pengetahuan kita sebagai manusia terhadap akar kata yang baik dan benar sangat diperlukan, karena semakin tinggi tingkat pengetahuan manusia terhadap akar kata maupun kebiasaan maka akan menghasilkan hasil yang lebih baik. Hal ini dilakukan untuk membandingkan akar kata hasil stemming menggunakan komputer dengan akar kata yang dihasilkan dari pengetahuan manusia.

Tabel 6. Penilaian relevansi pada dokumen

TIDAK	Kata Masukan	Hasil Steming (Akar Kata)	Penilaian Relevansi (Akar Kata)
1.	luar angkasa	luar angkasa	luar angkasa
2.	adalah	adalah	adalah
3.	atas	atas	atas
4.	atmosfer	atmosfer	atmosfer
5.	bulanan	bulan	bulan
6.	bumi	bumi	bumi
7.	dari	dari	dari
8.	lapisan	lapis	lapis
9.	gas	gas	gas
10.	yang	yang	yang

4.4. Mengukur Evaluasi

Algoritma stemming akan diuji menggunakan 100 dokumen. Adapun pengukuran yang dilakukan adalah sebagai berikut :

a) Hasil pengujian yang akurat diperoleh dari hasil perbandingan stemming dengan penilaian relevansi, dibagi dengan jumlah kata dalam dokumen.

- - - - - - - - - - × % (1)
- ! - -

b) Overstemming adalah kata yang banyak terpotong kata setelah proses stemming dibandingkan dengan penilaian relevansinya.

c) Understemming adalah kata-kata yang sedikit terpotong setelah melalui proses stemming dibandingkan dengan penilaian relevansi.

4.5. Hasil tes

Algoritma stemming akan diuji menggunakan 100 dokumen. Adapun pengukuran yang dilakukan adalah sebagai berikut: Hasil pengujian yang dilakukan terhadap nilai akurasi stemming proses menunjukkan bahwa nilai rata-rata akurasi terdapat pada algoritma stemming Porter.

Tabel 7. Hasil stemming yang akurat rata-rata

Nomor Dokumen (100)	Algoritma Porter
Akurasi rata-rata (%)	94.470

Hasil pengujian yang dilakukan terhadap nilai proses overstemming menunjukkan bahwa rata-rata overstemming pada algoritma stemming Porter.

Tabel 8. Persentase rata-rata overstemming

Nomor Dokumen (100)	Algoritma Porter
Rata-rata dari*berlebihan*(%)	4.541

Hasil pengujian yang dilakukan terhadap nilai proses understemming menunjukkan bahwa rata-rata understemming berada pada algoritma stemming Porter.

Tabel 9. Persentase rata-rata understemming

Nomor Dokumen (100)	Algoritma Porter
Rata-rata dari*meremehkan*(%)	0,989

Dalam algoritma Porter adalah ketika kata tersebut tidak ditemukan dalam database kamus dan kemudian dianggap sebagai akar kata. Terdapat kesalahan hasil stemming pada algoritma Porter terhadap imbuhan.

Tabel 10. Stemming menghasilkan kesalahan pada algoritma Porter.

Contoh	Hasil Steming	Seharusnya
Asupan	asupan	Asup
Bartahun	bartahun	Tahun
Bekerjasama	bekerja sama	Kerjasama
Beratnya	tikus	Berat
Berekpresi	berekpresi	Ekspresi
Berlaku	berla	laku
Berolah	Bero	Olah
berpengalaman	berpengalaman	Alam
Bersalah	bersa	Salah
Bersekolah	seko	Sekolah
sial	sial	Tanggungjawab
Bertanya	berta	Tanya
Bertopologi	bertopologi	Topologi
Berup	upa	Rupa
Bukanlah	bu	Bukan
Dariku	dari	Dari
Denganya	dengan	Dengan
Diadakan	adakan	Ada
penempatan	diinginkan	Anjur
Diataati	diataati	Taat

Referensi

[1] Agusta, Ledy (2009). Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia. Konferensi Nasional Sistem dan Informatika. KNS&I09-036.

[2] Tala, Fadillah. Z. 2004. Kajian Stemming Effect pada Information Retrieval dalam Bahasa Indonesia, Institute for Logic, Language and Computation Universiteit van Amsterdam Belanda

[3] Kamus Besar Bahasa Indonesia (KBBI) luring CHM V1.5, ebsoft.web.id. diakses Rabu 25 November 2015.

[4] Larkey, LS, Ballesteros, L., dan Connell, ME 2002. Meningkatkan Stemming untuk Pengambilan Informasi Bahasa Arab: Light Stemming dan Analisis Co-occurrence. Prosiding konferensi internasional tahunan ACM SIGIR ke-25 tentang Penelitian dan pengembangan pencarian informasi, 11-15 Agustus, Tampere, Finlandia.

[5] Nazief, Bobby dan Mirna Adriani. 1996. “Confix-Stripping: Pendekatan Algoritma Stemming untuk Bahasa Indonesia”. Fakultas Ilmu Komputer Universitas Indonesia. Nugraha, Lusianto

[6] Marga. 2010. Analisis Penggunaan Algoritma Stemming Vega pada Information Retrieval System. Universitas Telkom.Permana.

[7] Fanissa Shima, Ali Fauzi M., Adinugroho Sigit: Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.2018

[8] Hamzah Amir, Naniek : Opini Klasifikasi Menggunakan Entropi Maksimum dan K-Means Clustering, IEEE 2016

[9] Permana, A. Yudi, Ismasari Ismasari, dan M. Makmun Effendi. "Optimasi Stemming Porter KBBI dan Cross Validation Naïve Bayes untuk Klasifikasi Topik Soal UN Bahasa Indonesia." Jurnal Ilmiah KOMPUTASI 17.4 (2018): 357-368.

[10] Tan, A. 1999. Penambangan Teks: Kecanggihan dan tantangannya, Dalam Proc of the Pacific Asia Conf on Knowledge Discovery dan Data Mining Lokakarya PAKDD'99 tentang Penemuan Pengetahuan dari Database Tingkat Lanjut.